Development of The RU Hub4 system

C. Che, D. Yuk, S. Chennoukh, L. Flanagan

CAIP Center
Rutgers University
Piscataway, NJ 08855

ABSTRACT

This paper describes preliminary development of a broadcast news transcribing system for this year's Hub4 evaluation. The recognition system uses CROWNS (developed at RU for the 1995 Hub3 tasks) with several modifications to handle the news programming task. Features such as model adaptation have been added to quickly provide acoustic models thought appropriate for the new task, even though the environment-dependent data are limited. The architecture of decoding is changed from one pass to multi-pass that can handle higher order language models more efficiently. Due to the short development period before evaluation, the preliminary system for this year's Hub4 test has produced a higher error rate than expected. In fact, its performance is found to be worse than our previous system when compared on the baselin broadcast speech. We have continued investigation since the test and performed diagnostic experiments. Results and error analysis are given in this report.

1. INTRODUCTION

We report the development of the RU speech recognition system used in the 1996 Hub4 evaluation. We participated in the evaluation with the objective of building a recognition system for large vocabulary, continuous speech for the ARPA sponsored hands-free, distant-talking speech recognition project. This is also the first time we have experimented with the problem of transcribing broadcast news. Compared to last year's Hub3 test, the Hub4 task presents a more challenging problem with its great variety in speaking styles, channel conditions and presence of background noise and music.

We modified our existing system "CROWNS" [1] for this year's evaluation. To address the various focus conditions, features such as model adaptation are added to our system. The decoding strategy is also changed from last year's single-pass Viterbi beam search to a multi-pass word graph decoder that can efficiently handle higher order language models.

Due to lack of development time and experience, our system produces results with higher recognition word error rate than expected. It is concluded that this is a result of not performing enough experiments during development. A closer examination of the result from various focus conditions indicates that our system has problems dealing with the large vocabulary used in this year's Hub4 task. After the evaluation, several diagnostic experiments were performed. This paper will describe the experimental results and indicates directions for future work.

In the next section the preliminary system prepared for the 1996 Hub4 test is described including a new decoding scheme. Then we show the use of model adaptation to the Hub4 test. Diagnostic experiments and results are given in section 4. Finally, we present conclusions and future work directions.

2. SYSTEM DESCRIPTIONS

2.1. ACOUSTIC MODELS

CROWNS uses continuous density triphone HMM's as basic acoustic models. The HMM's are trained by conventional forward/backward (EM) algorithm. Each state of the triphone model use mixtures of Gaussians as output distributions. Transition probabilities are fixed. To obtain a more robust and reliable estimate for the huge number of Gaussian mixture used in the system, state tying are used with the furthest neighborhood tying suggested by [4].

There are 7 focus conditions for the primary tests in the Hub4 evaluation:

- *F0*: Baseline broadcast speech
- *F1*: Spontaneous broadcast speech
- *F2*: Speech over telephone channels
- *F3*: Speech in the presence of background music
- *F4*: Speech under degraded acoustic conditions
- *F5*: Speech from non-native speakers
- *FX*: All other speech

In recognition of the similarity of some focus conditions, four sets of basic acoustical models are constructed. The *WSJ SI284* corpus is used to bootstrap each set of the models. The conditions are described as follows:

- Full bandwidth SI284 (Set 1) to be used in *F0, F1, F5*
- Band-limited SI284 (Set 2) to be used in *F2*
- Full bandwidth SI284 plus music segments (Set 3) to be used in *F3*
- Secondary channel SI284 (Set 4) to be used in *F4*

Each set of models consists of 6930 word internal triphones with 8782 state clusters. Each state cluster contains 15 Gaussian mixtures with approximately 130K Gaussians in total. No gender models are used.

2.2. LANGUAGE MODEL AND LEXICON

With the currently available text materials, 3 sets of bigram/trigram language models (LM) are trained. The LM are generated using CMU SLM V1.0 with some fixes on processing the punctuation marks and removing extraneous words. Table 1 shows the trigram perplexity (PP) and OOV rate for 3 different LM's.

LM training corpus	PP (%)	OOV (%)
1995 Hub3 (LM1)	319.09	1.02
1996 Hub4 (LM2)	269.55	0.86
1995 Hub3 + 1996 Hub4 (LM3)	258.77	0.87

Table 1: 3gram perplexity (PP) and OOV rate for different LM training conditions

It is observed that by using the text from the 96 Hub4 only (LM2), we get about 15 % trigram perplexity reduction from LM1. Given the fact that the last condition produced lowest perplexity and almost the same OOV ratio to the 96 Hub4 training, our final system uses the LM trained with the combined corpus (LM3).

The recognition lexicon is obtained from the most frequent 60K words from LM3. Pronunciations of the 60k lexicon are extracted mainly from the CMU dictionary with approximately 600 words augmented by hand.

2.3. RECOGNITION

The recognition process is carried out in the following steps:

1. Acoustic segmentation: Input speech segments longer than 30 seconds are first broken into smaller chunks by using a energy based silence detector. The main reason for doing this extra segmentation is due to the memory constraints of our decoder.
2. Word graph construction: Viterbi bigram forward backward passes are then applied to the speech segments from step 1 and produce possible start/end time marks (boundaries). Word graph are then constructed from merging all possible word boundaries. The same acoustic models are used both in forward and backward search. Density of the graph is further reduced by pruning according to the best full path likelihood.
3. N Best rescoring: A final N best pass is used in producing alternative hypotheses. We use a word level A* algorithm [3] to search the word graph. The implementation is straight-forward since the graph already contains likelihood and boundaries for every word. Trigrams are used to rescore the upper 500 hypotheses, and output the top one as the result.

3. ADAPTATION

To make use of the available 60 hours worth of transcribed acoustic materials, a Maximum Likelihood Linear Regression transform based approach [2] is used. Due to the time constraint, only material from F0 and F3 conditions are used to adapt our seed model Set 1 and Set 3 with a global transform. The adapted models are then used in the final system decoding without performing any development test run.

4. POST EVALUATION ERROR ANALYSIS

The results of RU 96 Hub4 test are tabulated as follows

Average	F0	F1	F2	F3	F4	F5	FX
53.8	42.7	51.9	72.9	50.0	59.2	54.8	71.9

Table 2: Official word error rate (%) summary for the complete test set and focus condition

Compared to the results from other participating sites, our system produced relatively 30% to 40% higher word error rate. Certainly, this is not what we expected. After the evaluation, we have continued the investigation of analyzing errors and conducting several diagnostic experiments. From the results in table 2, it is clearly indicated that our system is not performing well for F0, the baseline broadcast speech. This condition has a dominant overal impact on all other conditions. The most obvious error we made is not performing any guiding experiments during the development period, even though the period was brief. We completed the model training and system integration only one week before the evaluation due to hardware problems. Model adaptation was used without any dry run on the Hub4 dev data. Those factors contributed to the poor performance. We have continued to perform several basic diagnostic experiments to answer the following questions:

4.1. Are our acoustic models appropriate?

One way to examine the question is to use this year's acoustic models in previous ARPA evaluation tasks. To simplify the recognition procedure, we run the 92 WSJ 5k and 20K tests using a standard bigram as LM. For comparison, a previously trained in-house model is used for the same test. Compared to the acoustic models we have in-house , this year system yields a 5 % absolute degradation. The only differences between the two systems are the tying parameters. This year's model is less tied and has more Gaussian mixtures. Originally, it is expected that a more relaxed tying will improve the performance, but on the contrary, it becomes the main degrading factor. Table 3 shows the comparison of recognition word error rates between the two systems.

From the table, it is found that more than 30% relative error reduction can be achieved by using different tying criteria. It is thus concluded that the acoustic models used for this year's test are certainly not good. We proceed to use the *better* acoustic models to decode the F0 portion data, and found the error rate reduced from 42.7% to 32% ! Obviously, the answer to the question is NO, and the acoustic models are certainly doing some damages to this year's evaluation.

System	#states/Gaussians	WSJ 5k	WSJ 20K
96 Hub4 models	8782/130K	16.0	20.2
In-house models	2333/80K	8.9	14.4

Table 3: Recognition word error rate (%) for different acoustic models

4.2. Is our LM appropriate?

We perform the F0 portion recognition using the LM2 and LM3. No difference in recognition word error rate is found. In terms of perplexity and OOV, these two are very similar. What is needed is to perform the same tests using the LM from 1995 Hub3 (LM1). It would be interesting to see the result and these experiment are under way.

4.3. Is adaptation helpful?

Since we had no test dry-run on the adaptation procedure, we suspect the adaptation lead to a system bug in producing this year's acoustic models. To verify our adaptation procedure, an environmental adaptation is performed in our hands-free distant speech recognition experiments. The goal is to adapt the models trained under clean speech to a more noisy and reverberant environment. Testing speech is recorded 12 ft away from the talker using a microphone array. We find that for the 1000 words recognition task (Distant-RM), the adaptation normally reduced the word error rate from 57% to 15%. Our best result for the same task is 9% using Neural Network based feature domain compensation. This result indicates the adaptation procedure itself has no problem.

The second experiment proceeds to run recognition using both adapted and seed models (Set 1) on the F0 portion data. No difference in recognition performance is found. No gain is observed from running the adaptation procedure. The answer to the question is NO, but not quite. It is still possible that we do not use enough transform (currently, only 1) for given amount of adaptation data. A larger number of transforms should be choosen from more experimental results.

5. CONCLUSIONS AND FUTURE WORK

In this report, the development of the 1996 RU Hub4 system is descibed. New features of this year's system include a modified multi-pass decoder and model adaptation procedure. Unfortunately, we have not gotten a stable system ready for this year's test. After the evaluation, several diagnostic experiments have been conducted in finding the sources of error. Here are some major observations:

- Need to run more experiments during development
- Main cause of error comes from acoustic models which are not robustly trained
- Not making full use of training data
- Need more understanding of the task

Realizing this year's recognizer produces higher error rate than the state of the art system, the focus of our future work is still on improving the basic unlimited vocabulary speaker independent speech recognition performance. In the post evaluation effort, we will continue to work on the various focus conditions and perform additional diagnostic experiments.

References

1. C. Che, *Development of CROWNS: CAIP Recognizer of Words N' Sentences*, DARPA Speech Recognition Work-Shop, Harriman, NY 1996, p. 112-116
2. C.J.Leggetter and P.C. Woodland, *Flexible Speaker adaptation Using Maximum Likelihood Linear Regression*, Proceedings of he Spoken Language System Technology Workshop, pp. 110-115, Jan. 1995
3. D. Paul and B. Necioglu, *The Lincoln Large-Vocabulary Stack-Decoder HMM CSR*, Proc ICASSP, Vol II, pp. 660-663, Minneapolis, 1993
4. P.C. Woodland and S.J. Young, *The HTK Tied State Continuous Speech Recognition*, Proc. EuroSpeech, Vol. 3, pp. 2207-2210, Berlin, 1993
5. Lee-K.F., *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, 1989.
6. H.Murveit, J.Butzberger, V.Digalakis and M. Weintraub, *Large-Vocabulary Dictation Using SRI's DECIPHER (TM) Speech Recognition System: Progressive-Search Techniques*, ICASSP, Vol II, pp. 319-322, Minneapolis, 1993
7. V.Steinbiss, B.-H.Tran and H.Ney, *Improvements in Beam Search*, Proc, Int.Conf. on Spoken Language Processing, pp. 2143-2146, Yokohama, Sep, 1994
8. S. Ortmanns, H.Ney, F.Seide and I.Lindam, *A Comparison of Time Conditioned and Word Conditioned Search Techniques for Large Vocabulary Speech Recognition*, Proc, Int.Conf. on Spoken Language Processing, Philadelphia, PA, Oct, 1996
9. G.Antoniol, F.Brugnara, M.Cettolo and M.Federico, *Language Model Representations for Beam Search Decoding*, Proc ICASSP, Vol. 1, pp. 588-591, Detroit, MI, May 1995
10. M.Oerder and H.Ney, *Word Graphs: An Efficient Interface Between Continuous Speech Recognition and Language Understanding*, Proc ICASSP, Vol. II, pp. 119-122, Minneapolis, MN, April, 1993